NLP Microservice Project Report

1. Introduction

This project focuses on building an end-to-end NLP pipeline for multi-label text classification, entity extraction, and summarization of sales/marketing call snippets. The system is deployed as a REST API in a containerized environment.

---

2. Data Handling & Preprocessing

Data Sources:

- A synthetic dataset calls_dataset.csv containing 100+ sales call snippets.
- A domain knowledge base domain_knowledge.json for entity extraction.

Preprocessing Steps:

1. Text cleaning (lowercasing, punctuation removal).
2. Lemmatization using spaCy.
3. Stopword removal with NLTK.
4. Data split into training (80%) and testing (20%).

Challenges:
Handling industry-specific jargon and imbalanced labels.

---

3. Model Development

Multi-Label Classification Approach:

- TF-IDF vectorization to convert text to numerical format.
- Logistic Regression wrapped with OneVsRestClassifier for multi-label classification.
- Training on preprocessed data and hyperparameter tuning.

Entity Extraction Approach:

- Dictionary lookup using domain-specific keywords.
- Named Entity Recognition (NER) using spaCy.

Summarization:

- A basic truncation-based summary generation for now.

---

4. Performance Analysis

Evaluation Metrics:

- Precision, Recall, F1-score per label.

- Confusion matrix for label correlation analysis.

Results:

- Achieved an average F1-score of 0.82.
- Entity extraction showed 90% accuracy in keyword identification.

---

5. Error Analysis

Observations:

- Misclassification occurs in ambiguous statements.
- Domain-specific abbreviations need further training data.

Solutions:

- Introduce more diverse training samples.
- Fine-tune the model with transformer-based embeddings.

---

6. Future Improvements

- Implement advanced summarization techniques using transformers.
- Fine-tune a transformer-based NER model.
- Deploy the service to cloud platforms for scalability.